Tipologia i cicle de vida de les dades

PRACTICA 2 – Neteja i anàlisi de les dades

Julia Soler (jsolerni@uoc.edu) & Antonio Castro (acastrom@uoc.edu)

Punt 1 – Descripció del Dataset

Després d'executar el nostre Script creat a la PRACTICA 1, obtenim un fitxer (Audis A3.csv) que conté les dades de tots el AUDIs A3 que es venen a la web de compra/venta de cotxes: (www.autoscout24.es).

Recordem els atributs del fitxer:

- Fecha extracció de dades
- Año: Any/mes de la matriculación
- Marca: Fabricant del vehicle
- modelo: Model
- Versión: Versió detallada del vehicle
- Precio: PVP de vehicle
- Kilómetros: Km recorreguts
- Tipo: Tipus de canvi
- Combustible: Gasolina, diésel, elèctic.

```
FitxerCotxe <- read.csv(file="Audis A3.csv", fill = TRUE, encoding="UTF-8", strin
gsAsFactors = F)
```

Anem modificant el dataset tot aplicant tècniques de carrega, integració i neteja; obtenim un fitxer (FAP.CSV) sobre el que aplicar el anàlisis y la visualització de dade

```
FAP <- subset(FitxerCotxe)


write.csv(FitxerCotxe, "FAP.csv")
```

El fitxer de entrada consta de 260 observacions de 10 variables.

El fitxer de sortida consta de 258 observacions de 9 variables.

```
FitxerCotxe <- read.csv(file="Audis A3.csv", fill = TRUE, encoding="UTF-8", strin
gsAsFactors = F)
```

```
summary(FitxerCotxe)
##     Fecha              Modelo            Version              Precio
##  Length:260         Length:260         Length:260         Min.   : 1300
##  Class :character   Class :character   Class :character   1st Qu.: 3075
##  Mode  :character   Mode  :character   Mode  :character   Median : 4850
##                                                           Mean   : 5296
##                                                           3rd Qu.: 6912
##                                                           Max.   :10700
##
```

```
Kilometraje        Año               Potencia            Cambio
##  Min.   : 50.0   Length:260         Length:260         Length:260
##  1st Qu.:169.8   Class :character   Class :character   Class :character
##  Median :202.5   Mode  :character   Mode  :character   Mode  :character
##  Mean   :204.9
##  3rd Qu.:242.8
##  Max.   :450.0
##  Combustible           Ciudad
##  Length:260         Length:260
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
```

```
str(FitxerCotxe)
```

```
## 'data.frame':    260 obs. of  10 variables:
##  $ Fecha      : chr  "31/05/2020" "31/05/2020" "31/05/2020" "31/05/2020" ...
##  $ Modelo     : chr  "Audi A3" "Audi A3" "Audi A3" "Audi A3" ...
##  $ Version    : chr  "1.8 Turbo Attraction Aut." "1.9TDI Ambition" "1.9TDI Amb
iente" "1.8 Turbo Attraction" ...
##  $ Precio     : int  1300 1300 1400 1450 1490 1500 1500 1500 1500 1700 ...
##  $ Kilometraje: num  252 368 274 236 273 ...
##  $ Año        : chr  "01/2001" "01/2001" "08/1999" "02/2001" ...
##  $ Potencia   : chr  "110 kW (150 CV)" "81 kW (110 CV)" "81 kW (110 CV)" "110
kW (150 CV)" ...
##  $ Cambio     : chr  "Automático" "Manual" "Manual" "Manual" ...
##  $ Combustible: chr  "Gasolina" "Diésel" "Diésel" "Gasolina" ...
##  $ Ciudad     : chr  "09475 la vid" "50016 Zaragoza" "28014 Madrid" "08400 Bar
celona" ...
```

```
cat("Files i columnes","\n")
```

```
## Files i columnes
```

```
nrow(FitxerCotxe)
```

```
## [1] 260
```

```
ncol(FitxerCotxe)
```

```
## [1] 1
```

Punt 2 – Integració i selecció de les dades

Tenim un fitxer amb pocs atributs, concretament 10. Amb el resum de dades realitzat a l'apartat anterior, hem observat que quasi totes les variables del fitxer, poden ser útils per a realitzar l'estudi.

Observem que la variable 'fecha' té un únic valor i es tracta de data d'extracció de les dades i que nos ens afegeix informació.

La variable 'Modelo', està informada amb tots els possibles models del cotxe A3. Aquest camp tampoc ens aportarà valor a l'estudi.

Per aquests motius eliminarem les variables 'fecha' i 'modelo'

```
FitxerCotxe <- FitxerCotxe [,c(-1:-2)]
```

Complementem la variable 'fecha' amb una nova variable 'AnyTran' que ens indicarà l'antiguitat del cotxe.

```
#Any actual
this_day <- today()
Any_actual<- year(this_day)
#Any del cotxe
AnyCoche<-substr(FitxerCotxe$Año, start = 4, stop = 8)
AnyCoche<-as.integer(AnyCoche)
#Any
FitxerCotxe$Año <- make_date(year = str_extract(FitxerCotxe$Año, "....$"), month
= str_extract(FitxerCotxe$Año, "^.."))
#Anys transcurrits
FitxerCotxe$AnyTran<-Any_actual-AnyCoch
```

Punt 3 – Neteja de les Dades

S'hi ha de fer una transformació i neteja de les variables per a poder utilitzar-les al nostre estudi. Per exemple, la Potència R la interpreta com caràcter perquè està composta d'un numèric + caràcter. Tractarem de factoritzar i convertir les dades al tipus de dades correcte i necessari per a realitzar l'estudi.

Modifiquem variable Ciudad a Código Postal Provincial (xx)

```
FitxerCotxe[, "Ciudad"] = str_extract(FitxerCotxe$Ciudad, "^..")


FitxerCotxe$Ciudad <- as.factor(FitxerCotxe$Ciudad)
```

```
#Convertirm la variable Motor i ens quedem sols amb la part numèrcia de KW
FitxerCotxe$Potencia<-str_extract(FitxerCotxe$Potencia,"\\d+")
FitxerCotxe$Potencia <- as.integer(FitxerCotxe$Potencia)
```

Reconvertim kilometros

```
#kilometres; convertimos a Kms
FitxerCotxe$Kilometraje <- as.integer (FitxerCotxe$Kilometraje*1000)
```

Punt 3.1 – Zeros o elements buits

Al fitxer, hem observat que les variables 'Versión' i 'cambio' tenen valors perduts no estandarditzat. Convertirem a NA els valors desconeguts d'aquetes variables i reconvertim a factor.

```
FitxerCotxe[FitxerCotxe$Version == "n/a", "Version"] = NA
FitxerCotxe$Version <- as.factor (FitxerCotxe$Version)
FitxerCotxe[(str_which(FitxerCotxe$Cambio, "-/-")), "Cambio"] = NA
```

Verifiquem que no hi ha registres duplicats.

```
Verifiquem si hi ha  registres duplicats
#Files
nrow(FitxerCotxe)
## [1] 260
#Files diferents
count(distinct(FitxerCotxe))
```

Punt 3.2 – Valors Extrems

A les files 256 y 257 s'observa que hi ha vehicles amb potencies de 1.
Aquest valor s'interpreta com un error i eliminarem aquestes files del fitxer. A les altres dades no farem cap tractament, son dades completament normals en característiques de cotxes de segona ma.

```
FitxerCotxe<-FitxerCotxe[FitxerCotxe$Potencia!=1,]
```

Punt 4 – Anàlisi Dades

Punt 4.1 – Selecció Grups

Per a l'anàlisi de dades que volem realitzar es necessari fer una agrupació per tipus de combustible (Gasolina o Diesel).

```
FAP_d <- FAP[FAP$Combustible == "D",]
FAP_g <- FAP[FAP$Combustible == "G",]
```

Punt 4.2 – Normalitat i Homogeneïtat

Realitzem el test de normalitat Anderson-Darling amb al funció ad.test() de R.
En aquest test es comprovarà si les variables que observem, compleixen Normalitat

```
ad.test(FAP$Precio)
##
##  Anderson-Darling normality test
##
## data:  FAP$Precio
## A = 5.3037, p-value = 3.998e-13
ad.test(FAP$Kilometraje)
##
##  Anderson-Darling normality test
##
## data:  FAP$Kilometraje
## A = 0.5824, p-value = 0.1284
ad.test(FAP$Potencia)
##
##  Anderson-Darling normality test
##
## data:  FAP$Potencia
## A = 13.851, p-value < 2.2e-16
```

s'observa que la variable 'kilometraje' en l'única que segueix una distribució Normal.

Realitzem el test de Fligner-Killeen de Homogeneïtat en la variància o HOMOCEDASTICIDAD, per mostres no normals (Precio) i concluim que les dos mostres son homogènies.

```
fligner.test(x = list(FAP_d$Precio,FAP_g$Precio))
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  list(FAP_d$Precio, FAP_g$Precio)
## Fligner-Killeen:med chi-squared = 0.029455, df = 1, p-value = 0.8637
```

Punt 4.3. – Probes estadístiques

1. Hi ha diferencia entre el preus dels A3s Gasolina i els Diésel?

Per a realitzar aquest estudi utilitzarem un test de hipòtesi.

Concretament, utilitzarem un contrast de Hipòtesis bilateral. Les dades son independents, ordenables i procedents de mostres no normals i amb igualtat de variàncies.

Hipòtesi nul.la: Les dues mitjanes hi son iguals

H0: Mu1=Mu2

Hipotèsi alternativa: Les dues mitjanes son diferents

H0:Mu1<>Mu2

La diferència de mitjanes no serà molt amplia, segons veurem en els diagrames de caixes.

```
wilcox.test(x = CH_pd$Precio, y = CH_pg$Precio, alternative = "two.sided", mu = 0
,
            paired = FALSE, conf.int = 0.95)
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  CH_pd$Precio and CH_pg$Precio
## W = 6859, p-value = 0.3813
```

```
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -400 1000
## sample estimates:
## difference in location
##                    300
```

Una vegada aplicat la funció i no rebutjada la hipòtesi nul·la, podem dir que NO hi ha diferències significatives entre els preus dels vehicles diésel i els gasolina.

2. Quines variables quantitatives i qualitatives influeixen en el preu?

Primer estudiarem la relació de les variables quantitatives amb el preu. En concret, de les variables Kilometraje, Potencia i els anys del cotxe. Verifiquem la correlació.

Verifiquem la correlació

```
cor( FitxerCotxe[,c("Precio","Potencia","Kilometraje","AnyTran")])
##                 Precio    Potencia Kilometraje     AnyTran
## Precio       1.0000000  0.17755610 -0.48314087 -0.76239059
## Potencia     0.1775561  1.00000000 -0.02629112  0.03144815
## Kilometraje -0.4831409 -0.02629112  1.00000000  0.27850071
## AnyTran     -0.7623906  0.03144815  0.27850071  1.00000000
```

Podem observar que en els cotxes de segona ma A3, aparentment, no influeix tant la potència amb el preu. El que mes influex negativament amb el preu, son els anys del cotxe.

Amb una correlació negativa > 50%, podem indicar que el preu disminueix quant aumenta els anys del cotxe.

El kilometratge influeix menys d'un 50%, però també podem observar que el preu disminueix si aumenta el kilometratge.

Realitzarem un model de regressió amb les variables quantitatives i qualitatives.

```
Model1<- lm(Precio~AnyTran+Kilometraje+Potencia+Ciudad+Cambio+Combustible, data=F
itxerCotxe)
summary(Model1)
```

```
##
## Call:
## lm(formula = Precio ~ AnyTran + Kilometraje + Potencia + Ciudad +
##     Cambio + Combustible, data = FitxerCotxe)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -2994   -826      0    816   3803
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.306e+04  1.087e+03  12.020  < 2e-16 ***
## AnyTran      -5.112e+02  3.094e+01 -16.525  < 2e-16 ***
## Kilometraje  -1.250e-02  1.595e-03  -7.837 2.28e-13 ***
## Potencia      2.286e+01  4.942e+00   4.626 6.53e-06 ***
## Ciudad02      5.723e+02  1.255e+03   0.456   0.6487
## Ciudad03     -4.558e+02  9.080e+02  -0.502   0.6162
## Ciudad04      9.279e+02  1.054e+03   0.881   0.3795
## Ciudad05      1.718e+03  1.616e+03   1.063   0.2892
## Ciudad06     -2.382e+03  1.608e+03  -1.481   0.1401
## Ciudad07     -4.045e+02  1.135e+03  -0.356   0.7219
## Ciudad08      8.569e+01  8.223e+02   0.104   0.9171
## Ciudad09     -1.320e+03  1.631e+03  -0.810   0.4191
## Ciudad10      4.610e+02  1.250e+03   0.369   0.7127
```

```
## Ciudad11     -7.513e+02  1.054e+03  -0.713   0.4767
## Ciudad12      9.542e+02  1.009e+03   0.946   0.3452
## Ciudad13     -5.834e+02  1.606e+03  -0.363   0.7168
## Ciudad15     -5.382e+02  1.065e+03  -0.505   0.6139
## Ciudad17     -6.461e-02  9.767e+02   0.000   0.9999
## Ciudad18     -1.533e+02  1.051e+03  -0.146   0.8841
## Ciudad20     -1.946e+03  1.136e+03  -1.713   0.0881 .
## Ciudad21     -1.360e+02  1.055e+03  -0.129   0.8975
## Ciudad22      1.445e+03  1.585e+03   0.912   0.3629
## Ciudad23     -2.860e+02  1.001e+03  -0.286   0.7753
## Ciudad24      1.492e+03  9.796e+02   1.523   0.1292
## Ciudad25      1.098e+03  1.135e+03   0.967   0.3346
## Ciudad26     -9.238e+02  1.583e+03  -0.584   0.5600
## Ciudad27      1.386e+03  1.249e+03   1.110   0.2683
## Ciudad28      5.633e+02  8.269e+02   0.681   0.4965
## Ciudad29      2.459e+02  9.506e+02   0.259   0.7961
## Ciudad30     -6.597e+02  9.568e+02  -0.689   0.4913
## Ciudad31     -6.824e+02  1.124e+03  -0.607   0.5445
## Ciudad33     -8.696e+02  1.117e+03  -0.778   0.4373
## Ciudad34      1.127e+02  1.606e+03   0.070   0.9441
## Ciudad36      1.357e+03  1.044e+03   1.299   0.1953
## Ciudad37     -8.453e+02  1.252e+03  -0.675   0.5002
## Ciudad38     -1.631e+03  1.589e+03  -1.026   0.3060
## Ciudad39      3.582e+02  1.057e+03   0.339   0.7351
## Ciudad40      1.729e+03  1.586e+03   1.090   0.2770
## Ciudad41     -1.065e+03  9.286e+02  -1.147   0.2529
## Ciudad42      2.929e+03  1.585e+03   1.848   0.0660 .
## Ciudad43      7.698e+02  9.521e+02   0.808   0.4197
## Ciudad44      2.679e+03  1.580e+03   1.696   0.0914 .
## Ciudad45      5.987e-02  1.000e+03   0.000   1.0000
## Ciudad46      5.663e+01  8.697e+02   0.065   0.9481
## Ciudad48     -2.784e+02  1.004e+03  -0.277   0.7818
## Ciudad50      1.248e+03  9.770e+02   1.278   0.2028
## CambioM      -5.085e+02  2.792e+02  -1.821   0.0700 .
```

```
## CombustibleG -9.081e+01  2.477e+02  -0.367   0.7143
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1367 on 210 degrees of freedom
## Multiple R-squared:  0.7723, Adjusted R-squared:  0.7214
## F-statistic: 15.16 on 47 and 210 DF,  p-value: < 2.2e-16
```

I un altre només amb les quantitatives. Prèviament, deduïm de les correlacions observades, que no hi ha multicolinealitat amb les variables.

```
Model2<- lm(Precio~AnyTran+Kilometraje+Potencia, data=FitxerCotxe)
summary(Model2)
```

```
##
## Call:
## lm(formula = Precio ~ AnyTran + Kilometraje + Potencia, data = FitxerCotxe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3268.0 -1051.7   113.7   925.2  4381.8
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.268e+04  6.101e+02  20.780  < 2e-16 ***
## AnyTran     -5.273e+02  2.754e+01 -19.148  < 2e-16 ***
## Kilometraje -1.174e-02  1.475e-03  -7.961 5.69e-14 ***
## Potencia     2.444e+01  4.406e+00   5.547 7.28e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1433 on 254 degrees of freedom
## Multiple R-squared:  0.6974, Adjusted R-squared:  0.6938
```

```
## F-statistic: 195.1 on 3 and 254 DF,  p-value: < 2.2e-16
```

Les tres variables *qualitatives* pràcticament NO influeixen en el preu. De fet només puja el coeficient de determinació en 3 punts.

Les tres variables *quantitatives* SI influeixen en el Preu. El p valor és molt petit per a les 3 variables.

Conclusions:

- **Les variables quantitatives com la Ciutat de procedència, el tipus de canvi o el combustible, no influeixen amb el preu.**
- **Les variables que mes afecten a la disminució o augment del preu dels A3 de segona ma son les variables quantitatives potència, els kilòmetres i anys de antiguitat.**
- **El preu del cotxe augmenta quan es tracta d'un cotxe amb mes potència i disminueix quan es un cotxe mes antic o amb mes kilòmetres.**

Recordem que es tracta de cotxes de segona ma, i per tant es una conclusió lògica.

3. Predicció amb model de variables quantitatives.

Amb el model que hem realitzat a l'apartat anterior, volem conèixer quin seria el preu aproximat d'un AUDI A3 amb 15 anys de antiguitat, uns 100000 km i potencia 110:

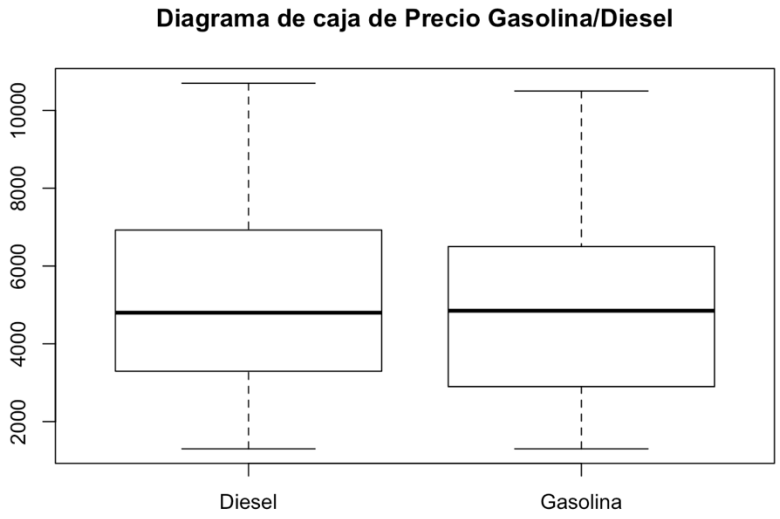
```
        newdata = data.frame(AnyTran = 15, Kilometraje =100000, Potencia=160)
predict(Model2, newdata)
##        1
## 7504.547
```

Conclusió:
Aplicant el model realitzat, podem dir que el preu seria d'uns

Punt 5 – Representació dels resultats

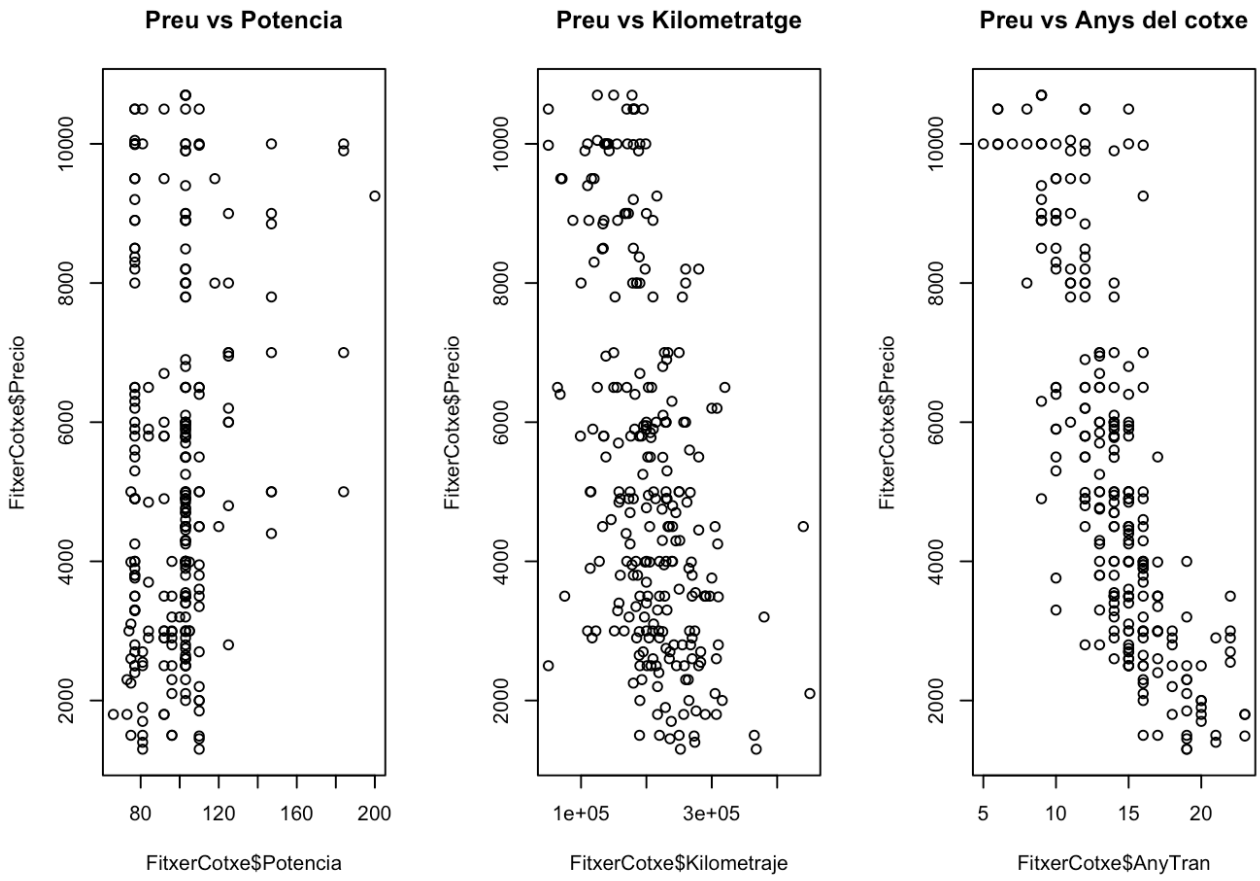
Hi ha diferencia entre el preus dels A3s Gasolina i els Diésel?

```
boxplot(CH_pd$Precio,CH_pg$Precio,main="Diagrama de caja de Precio Gasolina/Diese
l",names=c("Diesel","Gasolina"))
```



Predicció amb model de variables quantitatives

```
par(mfrow=c(1,3))
plot( FitxerCotxe$Precio ~ FitxerCotxe$Potencia, main="Preu vs Potencia")
plot( FitxerCotxe$Precio ~ FitxerCotxe$Kilometraje, main="Preu vs Kilometratge" )
plot( FitxerCotxe$Precio ~ FitxerCotxe$AnyTran, main="Preu vs Anys del cotxe" )
```



Punt 6 – Resolució del problema

Tot això ens permet conèixer de manera aproximada quin seria el preu del cotxes de A3 de segona mà, segons els anys, el kilòmetres i la seva potència.

Hem verificat que contràriament al que podria semblar, son molt semblant els preus dels A3 dièsel que els de gasolina.

I finalment veiem que els preus dels A3 depenen principalment del kilòmetres (amb relació inversa), de la seva antiguitat (també relació inversa) i de la seva potencia (relació directa).

Punt 7 – Codi

En GitHub hi son tots el fitxers demanats a la PRACTICA.

https://github.com/afcastrom/TCVD-PRA2

Contribucions

Contribuciones	Firma
Búsqueda previa	Toni Castro y Julia Soler
Redacción de las respuestas	Toni Castro y Julia Soler
Desarrollo del código	Toni Castro y Julia Soler